Instance Significance Guided Multiple Instance Boosting for Robust Visual Tracking

Jinwu Liu, Yao Lu, Tianfei Zhou

Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology

Abstract. Multiple Instance Learning (MIL) recently provides an appealing way to alleviate the drifting problem in visual tracking. Following the tracking-by-detection framework, an online MILBoost approach is developed that sequentially chooses weak classifiers by maximizing the bag likelihood. In this paper, we extend this idea towards incorporating the instance significance estimation into the online MILBoost framework. First, instead of treating all instances equally, with each instance we associate a significance-coefficient that represents its contribution to the bag likelihood. The coefficients are estimated by a simple Bayesian formula that jointly considers the predictions from several standard MILBoost classifiers. Next, we follow the online boosting framework, and propose a new criterion for the selection of weak classifiers. Experiments with challenging public datasets show that the proposed method outperforms both existing MIL based and boosting based trackers.

1 Introduction

Tracking-by-detection has emerged as a leading approach for accurate and robust visual tracking [2,3,6,8,13,14]. This is primarily because it treats tracking as a detection problem, thereby avoiding modeling object dynamics especially in the presence of abrupt motions [15] and occlusions [9]. Tracking-by-detection typically involves training a classifier to detect the target in individual frames. Once an initial detector is learned in the first frame, the detector will progressively evolve to account for the appearance variations in both the target and its surroundings.

It is well known that accurate selection of training samples for the detector updating is rather significant for a successful tracking-by-detection method. One common approach for this is to take the current tracking location as one positive example, and use the samples collected around the location for negatives. While this simple approach works well in some cases, the positive example used for detector updating may not be optimal if the tracking location is slightly inaccurate. Over time this will degrade the performance of the tracker. In contrast, many methods [2,6,8] use multiple positive examples for updating, where the examples are sampled from a small neighborhood around the current object location.

In principle, the latter updating scheme should be better because it exploits much more information. However, as reported in existing literature, it may confuse the appearance model since the label information about the positives is

not precise. Therefore, it may cause difficulties in finding an accurate decision boundary. Consequently, a suitable algorithm needs to handle such sort of ambiguities in training data, especially in the positive ones. Multiple Instance Learning (MIL) [5] can be exploited to achieve this goal, since it allows for a weaker form of supervision to learn with instance label uncertainty. For example, recent advances in object detection [1,12] demonstrate that MIL is able to largely improve the detection performance. Inspired by these applications, Babenko et al. [3] propose an online MILBoost approach to address the ambiguity problem in visual tracking. Along with this thread, Zhang et al. [13] propose an online weighted MIL tracker, and Bae et al. [4] introduce structural appearance representation into the MIL based tracking framework. In general, MIL enables these approaches to deal with slight appearance variations of the target during tracking, in which case, most instances in the positive bag are relatively close to a true positive. However, the trackers may fail in case of strong ambiguity, *e.g.*, motion blur, pose change, etc.

To address this gap, in this work, we follow the online boosting framework in [6] and propose a novel formulation of MILBoost for visual tracking. The central idea behind our approach is learning the significance of instances, which we call *significance-coefficients*, and incorporating them into the bag likelihood to guide the selection of weak classifiers in boosting. In particular, we begin by building a group of randomized MILBoost learners, and each provides its estimates for the instances being positive. Assuming that the learners are independent, we show that the significance-coefficients can be easily estimated through a simple Bayesian formulation. Further, we introduce a variant of bag likelihood function based upon the significance-coefficients for the selection of weak classifiers.

2 Proposed Approach

In the following, we first review the standard online multiple instance boosting method for tracking [3] and analyze its underlying limitations. This analysis motivates then our new extension, which allows for an accurate appearance model able to cope with diverse complex tracking scenarios.

2.1 Online Multiple Instance Boosting

Recently, Babenko et al. [3] propose a novel online boosting algorithm for MIL (online MILBoost) to address the example selection problem for adaptive appearance model updating in tracking. In particular, given a training data set $\{(X_1, y_1), \ldots, (X_n, y_n)\}$ in current frame, where a bag $X_i = \{x_{i1}, \ldots, x_{im}\}$ and $y_i \in \{0, 1\}$ is its label, as well as a pool of M candidate weak classifiers $\mathcal{H} = \{h_1, h_2, \ldots, h_M\}$, MILBoost sequentially chooses K weak classifiers from the candidate pool based upon the following criterion:

$$h_k = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \mathcal{L}(\mathbf{H}_{k-1} + h) \tag{1}$$

where $\mathcal{L} = \sum_i (y_i \log p_i + (1 - y_i) \log(1 - p_i))$ is the log-likelihood over bags, and $\mathbf{H}_{k-1} = \sum_{i=1}^{k-1} h_i$ is the strong classifier consists of the first $k-1$ weak classifiers. Note that $\mathcal{L}$ is the bag likelihood rather than instance likelihood used in traditional supervised learning approaches, and p_i indicates the probability of bag i being positive, which is defined by the Noisy-OR model:

$$p_i = p(y_i|X_i) = 1 - \prod_j (1 - p(y_i|x_{ij})) \tag{2}$$

and $p(y_i|x_{ij}) = \sigma(\mathbf{H}(x_{ij}))$ is the instance probability where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

Note that the Noisy-OR model in Eq. 2, which is used to account for ambiguity, holds an assumption that all instances in a bag contribute equally to the bag likelihood. It is imprecise because according to the MIL formulation, a positive bag contains at least one positive instance, but it may also contain many negative ones. Clearly, the model in Eq. 2 cannot identify the true positives in the positive bags. While [13] mitigates this problem using a weighted MILBoost method, we observe that slight inaccuracies in tracking results will lead to inaccurate weights, thereby degrading the tracking performance. Furthermore, not only is the likelihood model too restrictive, but also one single MILBoost is not flexible enough for capturing the multi-modal distribution of the target appearance.

2.2 Significance-Coefficients Estimation

The previous analysis motivates our extension of standard MILBoost to a more robust model so that it can handle various challenging situations. Here we aim to integrate the instance significance into the learning procedure. Note that our method is essentially different from [13] because we in this work determine the instance significance discriminately rather than simply weighting the instances according to Euclidean distances between the instances and the object location.

In particular, we begin with training N learners:

$$\Phi = \{\mathbf{H}_1, \ldots, \mathbf{H}_N\} \tag{3}$$

where $\mathbf{H}_i$ denotes a randomized MILBoost classifier learned in Sec. 2.1, and the randomization is obtained by sampling different negative examples for each learner. Then, for each instance x_{ij}, its significance-coefficient r_{ij} is jointly determined by the predictions of the learners:

$$r_{ij} = p(y_{ij}|\mathbf{H}_1, \ldots, \mathbf{H}_N) \tag{4}$$

where y_{ij} denotes the label of x_{ij}. Assuming that the randomized MILBoost classifiers are conditional independent, we can rewrite the above formulation as:

$$r_{ij} \propto p(\mathbf{H}_1, \ldots, \mathbf{H}_N|y_{ij}) p(y_{ij}) \tag{5}$$

$$\propto p(y_{ij}) \prod_{k=1}^{N} p(\mathbf{H}_k|y_{ij}) \tag{6}$$

Note that we also have $p(\mathbf{H}_k|y_{ij}) = \frac{p(y_{ij}|\mathbf{H}_k)p(\mathbf{H}_k)}{p(y_{ij})}$, then the above formulation is equivalent to:

$$r_{ij} \propto p(y_{ij}) \prod_{k=1}^{N} \frac{p(y_{ij}|\mathbf{H}_k)}{p(y_{ij})} \tag{7}$$

where $p(y_{ij})$ is the prior indicating the probability that x_{ij} is positive, *i.e.*, $y_{ij} = 1$, and $p(y_{ij}|\mathbf{H}_k) = \sigma(\mathbf{H}_k(x_{ij}))$ is the prediction of $\mathbf{H}_k$ over instance x_{ij}.

Eq. 7 has two characteristics in computing the significance-coefficients: 1) if the predicted probability $p(y_{ij}|\mathbf{H}_k)$ is larger than the prior $p(y_{ij})$, the significance of x_{ij} will be enhanced; 2) considering the multiplicative part $\prod_{k=1}^{N} p(y_{ij}|\mathbf{H}_k)$, each predicted value can be viewed as imposing a weight to other predictions. This intuitively benefits the significance estimation procedure.

Given the significance-coefficients of all instances in a positive bag, we follow the underlying philosophy of MIL to estimate the bag significance:

$$r_i = \max_j r_{ij} \tag{8}$$

It should be noted that in MIL, ambiguity only exists in the positive bags. Hence, we only estimate the significance-coefficients for instances in the positive bags, but fix the significance of negative instances to $r_{ij} = 1$, thus $r_i = 1$.

2.3 Refinement of Online MILBoost

As introduced before, the Noisy-OR model is not precise because it does not take the instance significance into account. In this work, we extend the Noisy-OR model in Eq. 2 to the following:

$$p_i = p(y_i|X_i) = 1 - \prod_j (1 - p(y_i|x_{ij}))^{\alpha \frac{r_{ij}}{r_i}} \tag{9}$$

The novel exponent term enables us to integrate the instance significance into Eq. 2. In particular, the instance x_{ij} is equivalent to repeat $\alpha \frac{r_{ij}}{r_i}$ times in the bags, and α is a constant that denotes the possible maximal repetition number for the instances. In fact, in our experiments, we set $\alpha = 1$ for the negative bags so that Eq. 9 is equivalent to Eq. 2, and empirically set $\alpha = 3$ for the positive bags to incorporate instance significance.

Next, we develop an extended log-likelihood function over the bags as:

$$\mathcal{L}_e = \sum_i r_i (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{10}$$

Given the new log-likelihood function, we train a boosted classifier of weak learners as in [3]:

$$h_k = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \mathcal{L}_e(\mathbf{H}_{e,k-1} + h) \tag{11}$$

This is similar to the procedure in Eq. 1, except that we use a novel likelihood function $\mathcal{L}_e$ instead of $\mathcal{L}$ for weak classifier selection. Finally, we obtain a strong classifier $\mathbf{H}_e$ used as our discriminant appearance model.

2.4 Weak Classifiers

In this work, each object bounding box is represented using a set of Haar-like features [10]. Each feature consists of 2 to 4 rectangles, and each rectangle has a real-valued weight. Thus, the feature value is a weighted sum of the pixels in all the rectangles.

For each Haar-like feature f_k, we associate it with a weak classifier h_k with four parameters $(\mu_1, \sigma_1, \mu_0, \sigma_0)$:

$$h_k(x) = \log \frac{p_t(y=1|f_k(x))}{p_t(y=0|f_k(x))} \tag{12}$$

where $p_t(f_t(x)|y=1) \sim \mathcal{N}(\mu_1, \sigma_1)$ and similarly for $y=0$. Note that the above equation establishes with a uniform prior assumption, *i.e.*, $p(y=1) = p(y=0)$.

Following [3], we update all the weak classifiers in parallel when new examples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ are passed in:

$$\mu_1 \leftarrow \gamma\mu_1 + (1-\gamma)\frac{1}{n}\sum_{i|y_i=1} f_k(x_i) \tag{13}$$

$$\sigma_1 \leftarrow \gamma\sigma_1 + (1-\gamma)\sqrt{\frac{1}{n}\sum_{i|y_i=1}(f_k(x_i)-\mu_1)^2)} \tag{14}$$

where $\gamma \in [0,1]$ is the learning rate. The update rules for μ_0 and σ_0 are similarly defined. Note that our randomized MILBoost learners Φ and the new classifier $\mathbf{H}_e$ share the pool of candidate weak classifiers, as well as the updating rules.

2.5 Tracking Algorithm

In this section, we summarize our tracking algorithm. Without loss of generality, we assume the object location ℓ^*_{t-1} at time $t-1$ is given. 1) We first crop out some image patches $X^\gamma = \{x : \| \ell(x) - \ell^*_{t-1} \| < \gamma\}$ as the positive instances, and other ones $X^\beta = \{x : \gamma < \| \ell(x) - \ell^*_{t-1} \| < \beta\}$ as the negative instances, where $\ell(x)$ denotes the location of patch x , γ and β are two scalar radius (measured in pixels). 2) Given the training examples, we learn a group of randomized MILBoost classifiers Φ as well as an improved MILBoost classifier $\mathbf{H}_e$. 3) At time t, we crop out a set of image patches $X^s = \{x : \| \ell(x) - \ell^*_{t-1} \| < s\}$ where s is a small search radius. 4) The object location ℓ^*_t is ultimately obtained by:

$$\ell^*_t = \ell\left(\underset{x \in X^s}{\operatorname{argmax}}\, p(y|x)\right) \tag{15}$$

where $p(y|x) = \sigma(\mathbf{H}_e(x))$ is the appearance model. For other frames, our tracker repeats the above procedure to capture the object locations.

Fig. 1. Illustration results of the Boy sequence. (Best viewed in color)

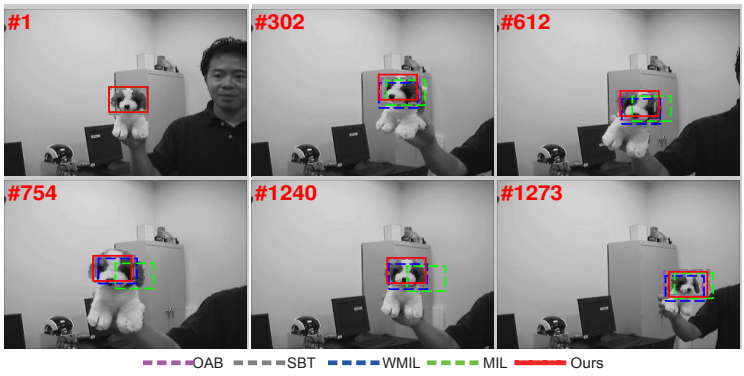

Fig. 2. Illustration results of the Dog sequence. (Best viewed in color)

3 Experiments

To evaluate the performance of the proposed algorithm thoroughly, we perform experiments on nine publicly available sequences with different challenging properties. The total number of frames we tested is more than 9000. We compare the method against other 4 state-of-the-art algorithms: MIL [3], WMIL [13], OAB [6], and SBT [7]. For fair comparison, we run the source codes provided by the authors with tuned parameters to obtain their best performance.

Our tracker is implemented in MATLAB and runs at 15 frames per second on a 2.93GHz Intel Core i7 CPU. In the experiments, the search radius s is set to 25 pixels, and the scalars γ and β are set to 4 and 50 respectively. For the negative image patches, we randomly select 200 patches from X^{β}. Then, Φ and $\mathbf{H}_e$ are online updated using only 50 of 200 negative patches. The number of randomized MILBoost classifiers is set to $\| \Phi \| = 3$, and the learning rate in

Table 1. Average Center Location Error (in pixel). Top two results are shown in **Red** and **Blue** fonts.

Seq	OAB	SBT	WMIL	MIL	Ours
Boy	**6.8**	7.3	58.1	30.3	**5.4**
Dog	19.9	27.8	**13.3**	25.5	**10.2**
Doll	19.6	**12.1**	41.5	30.2	**8.3**
Dollar	38.2	77.6	**35.1**	74.3	**9.3**
Girl	25.0	**18.0**	54.4	38.8	**20.2**
Panda	8.2	7.2	**6.3**	7.8	**6.7**
Sylv	18.7	**17.0**	19.9	44.7	**14.5**
Twinings	33.9	**19.7**	21.7	20.5	**18.3**
Walking	**5.2**	5.3	11.9	6.6	**5.0**
Average	**19.5**	21.3	29.1	31.0	**10.9**

Table 2. Average Overlap Rate. Top two results are shown in **Red** and **Blue**.

Seq	OAB	SBT	WMIL	MIL	Ours
Boy	**0.67**	0.53	0.43	0.48	**0.78**
Dog	0.40	0.39	0.45	**0.47**	**0.48**
Doll	0.59	**0.64**	0.39	0.34	**0.75**
Dollar	0.61	0.25	**0.63**	0.29	**0.73**
Girl	0.54	**0.71**	0.44	0.41	**0.62**
Panda	0.73	**0.80**	0.71	0.76	**0.79**
Sylv	**0.65**	0.64	0.60	0.43	**0.70**
Twinings	0.54	**0.81**	0.55	**0.57**	**0.81**
Walking	**0.71**	0.70	0.51	0.64	**0.74**
Average	0.60	**0.61**	0.52	0.49	**0.71**

Eq. 13 and Eq. 14 is fixed to $\gamma = 0.85$. Finally, the number of weak classifiers M is set to 150, and each time $K = 15$ classifiers are chosen to form a strong classifier.

3.1 Quantitative Evaluation

We employ two widely used evaluation criteria to evaluate the performance of the trackers: 1) *Center Location Error(CLE)* which measures the position errors between central locations of the tracking results and the centers of the ground truth; 2) *VOC Overlap Rate (VOR)* that evaluates the success ratio of the algorithms, which is calculated by $VOR = \frac{|B_r \cap B_g|}{|B_r \cup B_g|}$, where B_r denotes the tracked bounding box, B_g is the ground truth box, and $|\cdot|$ denotes the number of pixels in a region. Tab. 1 and Tab. 2 respectively summarize the average CLEs and

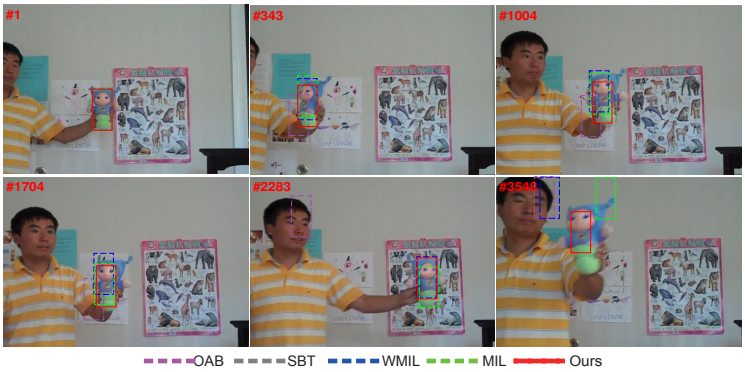

Fig. 3. Illustration results of the Doll sequence. (Best viewed in color)

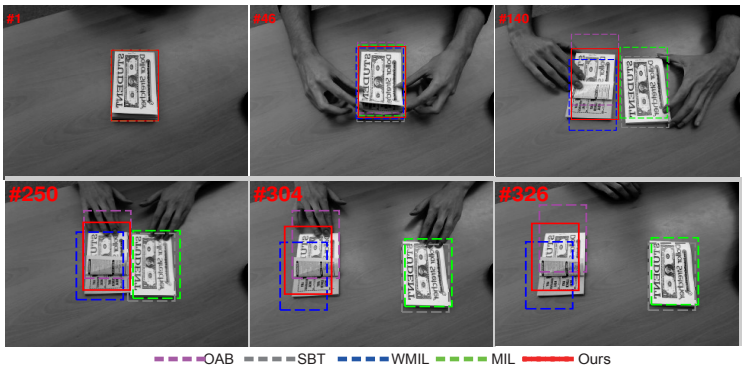

Fig. 4. Illustration results of the Dollar sequence. (Best viewed in color)

the average VORs of the compared trackers on the nine videos. The potential benefits of our tracker are notable: it performs best on 7 of 9 videos in terms of the average CLEs as well as the average VORs. Compared with MIL and WMIL, the performance improvement is particularly impressive in *Boy*, *Doll* and *Walking* sequences. As discussed in [11], these sequences, which contain *motion blur* and *low-resolution target*, are highly challenging for previous MIL based trackers. In our study, the multiple randomized classifiers enable us to capture the complex multi-modal distribution of the target appearance. Furthermore, our bag likelihood function is more accurate than the previous algorithms. Hence, our algorithm can better handle these challenges.

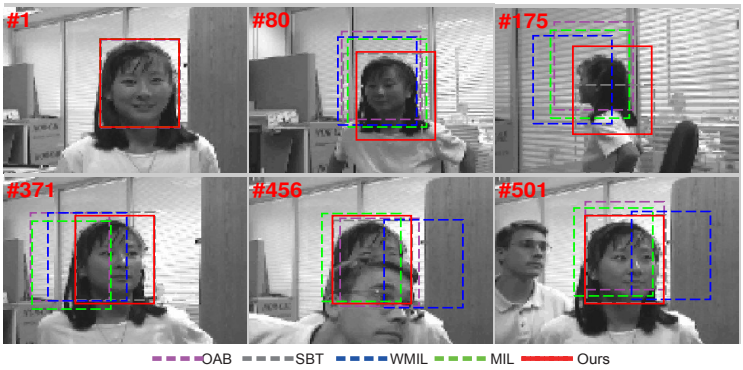

Fig. 5. Illustration results of the Girl sequence. (Best viewed in color)

Fig. 6. Illustration results of the Panda sequence. (Best viewed in color)

3.2 Qualitative Evaluation

In this section, we qualitatively compare our method with other trackers in dealing with various challenging factors.

Fast Motion: We firstly evaluate these trackers on two sequences with fast motion, which are *Boy* and *Doll*. These sequences are challenging because fast movement may result in blurred object appearance that is difficult to handle in object tracking. As shown in Fig. 1, MIL and WMIL fail to track the target before frame #94 in *Boy* in which the target appearance undergoes significant change, and OAB and SBT also locate the target inaccuratly in frame #227 and #541. Our method can track the sequences successfully with a small error mainly because of the more accurate likelihood function.

For the *Doll* sequence, object appearance changes drastically as the target moving back and forth. As illustrated in Fig. 3, OAB easily loses the target at

the beginning of the sequences (*e.g.*, #343). Subsequently, SBT, WMIL, and MIL also does not deal with the motion blur well when the target undergoes fast motion (*e.g.*, #3548). Overall, our algorithm can accurately estimate the location of the target throughout the sequence.

Pose Change: We next evaluate our method in dealing with larget pose change over four challenging sequences, *i.e.*, *Sylv*, *Girl*, *Dog* and *Twinings*. In both *Sylv* and *Dog* sequences, the targets suffer from great pose change. As show in Fig. 7 and Fig. 2, the previous multiple instance learing based trackers, *i.e.*, MIL and WMIL fail in these situations since the likelihood of the target is not accurately estimated. After a long-term tracking, the two trackers generally lose the target (*e.g.*, #884, #1171, and #1273 in *Sylv*, #1240 in *Dog*, #70, #165, #188, #224). In contrast, OAB, SBT and our tracker can well handle the appearance change caused by pose change and give better results.

For the *Girl* and *Twinings* sequences, the objects suffer from out-of-plane rotations as well as heavy occlusion. The WMIL algorithm performs worst in these two sequences, as shown in Fig. 5 and Fig. 8. OAB, SBT and our tracker are able to track the target in the two sequences.

Other Challenges: As shown in Fig. 6 and Fig. 9, the *Panda* and *Walking* sequences show that our method copes well with the situations where the target is actually of low-resolution, primarily because our method can select more discriminative features in the boosting stage than the previous approaches.

For the *Dollar* sequence, there is a distractor which may result in the failure of the trackers. As presented in Fig. 4, the MIL and the SBT trackers easily drift from the target due to the object with similar appearance. The OAB, WMIL and our tracker perform well in this challenging situtation. Besides, our algorithms gives more accurate tracking results than the two methods, as illustrated in Tab. 1 and Tab. 2.

Finally, it's revealed in *Dog* and *Doll* sequences that our tracker is more stable than other methods during long-term tracking, owing that by incorporating significance-coefficients of instances, our MIL method can well handle the ambiguity when updating the appearance model, as shown in Fig. 2 and Fig. 3.

4 Conclusion

Inspired from the recent success of multiple instance learning (MIL) in tracking, we proposed a novel algorithm that incorporates the significance-coefficients of instances into the online MILBoost framework. Our approach consists of two steps: (i) significance-coefficients estimation via a Bayesian formulation based on the predictions given by the randomized MILBoost classifiers, and (ii) a flexible scheme for incorporating the instance significance into the objective function of online MILBoost. In the experiments, we evaluate our method on several publicly available datasets and the results show its better performance.

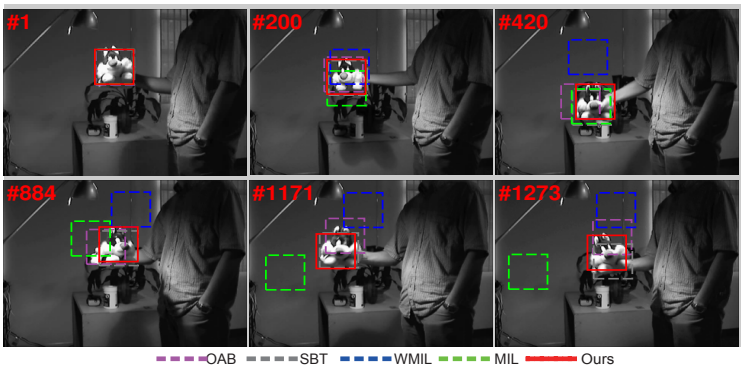

Fig. 7. Illustration results of the Sylv sequence. (Best viewed in color)

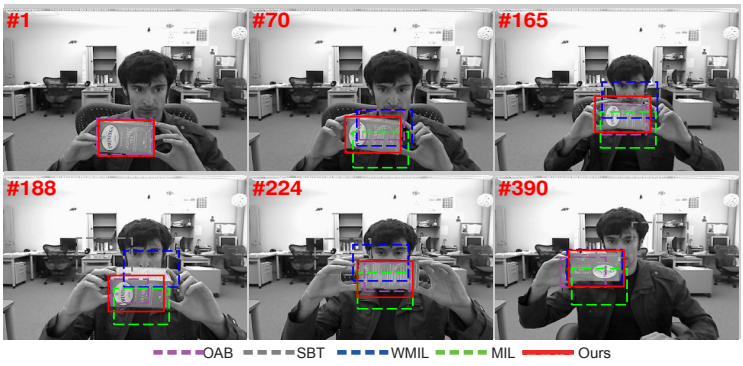

Fig. 8. Illustration results of the Twinings sequence. (Best viewed in color)

References

1. Ali, K., Saenko, K.: Confidence-rated multiple instance boosting for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2433–2440 (June 2014)
2. Avidan, S.: Ensemble tracking. IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI) 29(2), 261–271 (2007)
3. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI) 33(8), 1619–1632 (2011)
4. Bae, S.H., Kim, M.: Object tracking based on online partial instance learning with multiple local strong classifiers. In: IEEE International Conference on Image Processing (ICIP) (2014)
5. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence 89(1), 31–71 (1997)

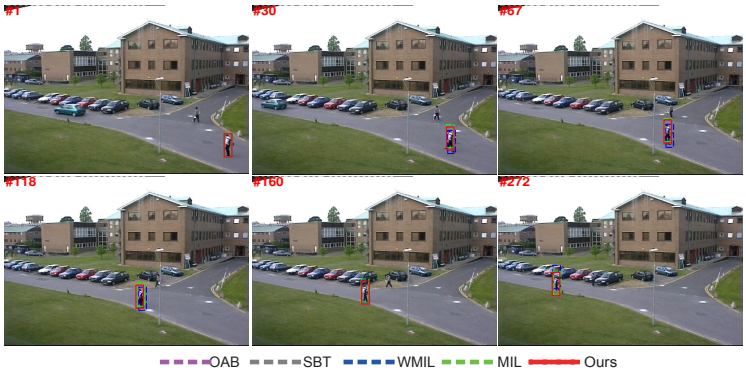

Fig. 9. Illustration results of the Walking sequence. (Best viewed in color)

6. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: BMVC. vol. 1, p. 6 (2006)
7. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: European Conference on Computer Vision (ECCV), pp. 234–247. Springer (2008)
8. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI) 34(7), 1409–1422 (2012)
9. Pan, J., Hu, B.: Robust occlusion handling in object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8 (2007)
10. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. I–511 (2001)
11. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Recognition and Machine Intelligence (TPAMI) (2015)
12. Zhang, C., Platt, J.C., Viola, P.A.: Multiple instance boosting for object detection. In: Advances in Neural Information Processing Systems. pp. 1417–1424 (2005)
13. Zhang, K., Song, H.: Real-time visual tracking via online weighted multiple instance learning. Pattern Recognition 46(1), 397–411 (2013)
14. Zhou, Q.H., Lu, H., Yang, M.H.: Online multiple support instance tracking. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. pp. 545–552. IEEE (2011)
15. Zhou, T., Lu, Y., Di, H.: Nearest neighbor field driven stochastic sampling for abrupt motion tracking. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2014)